SciDAC's Earth System Grid Center for Enabling Technologies
Semi-Annual Progress Report for the Period
October 1, 2009 through March 31, 2010

Principal Investigators
Dean N. Williams[3], Ian T. Foster[1], and Don E. Middleton[5]

The Earth System Grid Center for Enabling Technologies Team:
Rachana Ananthakrishnan[1], Frank Siebenlist[1]
Arie Shoshani[2], Alex Sim[2,]
Gavin Bell[3], Robert Drach[3]
Jim Ahrens[4], Phil Jones[4]
David Brown[5], Julien Chastang[5], Luca Cinquini[5], Peter Fox[5], Danielle Harper[5], Nathan Hook[5], Eric Nienhouse[5], Gary Strand[5], Patrick West[5], Hannah Wilcox[5], Nathaniel Wilhelmi[5], Stephan Zednik[5]
Steve Hankin[6], Roland Schweitzer[6]
Meili Chen[7], Ross Miller[7], Galen Shipman[7], Feiyi, Wang[7]
S. Bharathi[8], Ann Chervenak[8], Robert Schuler[8], Mei-Hui Su[8]

Climate simulation data are now securely accessed, monitored, cataloged, transported, and distributed to the national and international climate community

[1] Argonne National Laboratory
[2] Lawrence Berkeley National Laboratory
[3] Lawrence Livermore National Laboratory
[4] Los Alamos National Laboratory
[5] National Center for Atmospheric Research
[6] National Oceanic and Atmospheric Administration
[7] Oak Ridge National Laboratory
[8] University of Southern California, Information Sciences Institute

Table of Contents

1   Executive Summary

This report summarizes work carried out by the Earth System Grid Center for Enabling Technologies (ESG-CET) during the period October 1, 2009 through March 31, 2010. It includes discussion of highlights, overall progress, period goals, collaborations, papers, and presentations. To learn more about our project, and to find previous reports, please visit the ESG-CET website (URL: http://esg-pcmdi.llnl.gov/). This report will be forwarded to the Department of Energy (DOE) Scientific Discovery through Advanced Computing (SciDAC) program management, the Office of Biological and Environmental Research (OBER) program management, national and international collaborators and stakeholders (e.g., the Coupled Model Intercomparison Project, phase 5 (CMIP-5) for the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report (AR5), the Community Climate System Model (CCSM), the Climate Science Computational End Station (CCES), the SciDAC II: A Scalable and Extensible Earth System Model for Climate Change Science, the North American Regional Climate Change Assessment Program (NARCCAP), and other wide-ranging climate model evaluation activities).

The ESG-CET executive committee consists of Dean N. Williams Lawrence Livermore National Laboratory (LLNL); Ian Foster, Argonne National Laboratory (ANL); and Don Middleton, National Center for Atmospheric Research (NCAR). The ESG-CET team is a group of researchers and scientists with diverse domain knowledge, whose home institutions include seven laboratories and one university: ANL, Los Alamos National Laboratory (LANL), Lawrence Berkeley National Laboratory (LBNL), Lawrence Livermore National Laboratory (LLNL), NCAR, Oak Ridge National Laboratory (ORNL), Pacific Marine Environmental Laboratory (PMEL), and University of Southern California, Information Sciences Institute (USC/ISI). All work is accomplished under the DOE open-source guidelines and in close collaboration with the project's stakeholders, domain researchers, and scientists.

Overall goal for this reporting period

Our overall goal of this reporting period was to prepare for receiving and disseminating the CMIP-5 (IPCC AR5) data.

Highlights

1.1.1   *ESG Release Schedule*

ESG has become a very important project for a number of climate initiatives (e.g., CMIP-5, CCSM, NARCCAP – more listed above), so software new release and update must be given due diligence. To be able to make a release of ESG, cooperation between all the developers and all the collaborating sites is needed. A schedule was made in an effort to clarify to everyone on the team and to the community what is required to make a successful release. This schedule is viewed more as a guideline and not as a strict plan for releasing the software, which consist of disparate software making up the Gateway and the Data Node.

1.0.0-BETA 2	Goal: Limited to ESG collaboration for testing, review, and evaluation	Date: Wednesday, February 17, 2010
- Blocker and critical issues and features for the first production release. - OpenID-based Single Sign-ON (SSO). - Demonstrate prototype Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server (TDS) security with OpenID and Secure Sockets Layer (SSL). - Get core versioning capabilities into the data model, with existing user interface (UI) functioning properly on top (UI additions will come later). - Demonstrate Curator/METAFOR enhancements to system. - Demonstrate prototype support for Data Reference Syntax (DRS) in terms of publication, browse, and search. - Begin supporting ORNL as part of the federation. - Engage European partners in exercising and evaluating. - Upgraded downloads to incorporate versioning.		
1.0.0-RC1	Goal: Limited announcement to climate scientist for evaluation	Date: Tuesday, March 2, 2010
- No further database re-initialization (publisher and gateway). - Functionally complete for release, blocker and critical items for the first production release. - Generation of reusable (token less) WGET scripts. - Generation of Wget scripts for multiple data sets for a single Gateway. - Support for new Data Mover-Light (DML) scripts. - Transfer existing users. - Improve download user interface responsiveness. - Incorporate authorization service. - Complete publishing existing ESG CMIP3 data holdings.		
1.0.0-RC2 – RC5	Goal: Limited announcement to climate scientist for evaluation	Date: Monday, March 8, 2010 through Monday, April 19, 2010
Re-enable previous levels of hard delete functionality.		
1.0.0-Official Release	Goal: Enable federation services	Date: Monday, May 3, 2010
- OpenID login. - Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) harvesting.		
1.1-Minor Release	Goal: Additional versioning support and security infrastructure	Date: Monday, June 1, 2010
- Removal of tokens, switch to new security infrastructure. - Support Data Node metrics and notification with gateway-side service for retrieval of user information. - Fully support deletion of data sets. - Verify database integrity after hard deletion. - Support logical deletion. - Managing file access points when versioning.		
1.2-Minor Release	Goal: Open to the community for evaluation	Date: Monday, June 28, 2010
- Enable better data download workflows (via Data Cart capabilities). - Live Access Server (LAS) integration.		
1.3-Minor Release	Goal: Support CMIP-5 distributed archive	Date: Monday, August 2, 2010
- Data access by Designated Record Set (DRS) Websites (URLs). - Support replication. - Store necessary replica metadata in gateway database. - Search of replicated data sets. - Versioning UI.		

Table 1: Source code release schedule of the ESG enterprise distributed system.

1.1.2   *SC'09 Bandwidth Challenge*

The CMIP-3 data set consists of many small-sized files spanning spatial and temporal time zones making the "multiple small file" transfer optimizations in the Grid File Transfer Protocol (GridFTP), such as data channel caching and pipelining, relevant to the efficiency of the data transfer protocol. The Bulk Data Mover (BDM), higher-level data transfer management component was used to manage the GridFTP transfers with optimized transfer queue and concurrency management algorithms. The network bandwidth reservation capabilities provided by the Energy Sciences Network (ESnet) On-demand

Secure Circuits and Advance Reservation System (OSCARS) utilized dedicated circuits for the data transfers.

This Bandwidth Challenge activity, led by ANL with participations from ANL's Argonne Leadership Computing Facility (ALCF), ESnet, LBNL, LLNL, the National Energy Research Scientific Computing Center (NERSC), and the University of Utah was particularly important for ESG and the climate community to prepare global data replication over the network for the upcoming CMIP-5 (IPCC AR5) release. It showed our current status in the data replication over the network and what must be done to make transfers work better.

Data source was setup at ALCF, NERSC, and LLNL. The size of the data set was 10 terabytes (TB), and the transfers were partitioned into three groups with 4 TB from ALCF, 4 TB from NERSC, and 2 TB from LLNL. Network bandwidth on the Supercomputing 2009 (SC'09) showroom floor was 20-gigabits-per-second (Gbps). ESnet Science Data Network (SDN) was reserved for the transfers through ESnet OSCARS at 10-Gbps from ALCF, 10-Gbps from NERSC, and 5-Gbps from LLNL. The reason that data set was partitioned into this setup was because 4 TB is the volume of the file transfers achievable in an hour over a 10-Gbps connection. When network bandwidth was fully utilized under the setup, 10 TB could be transferred in about 1.5 hours. **Figure 1** shows how the network was setup and how data flowed.

Figure 1: SC'09 data and network setup diagram (image courtesy Dan Gunter, LBNL and Raj Kettimuthu, ANL).
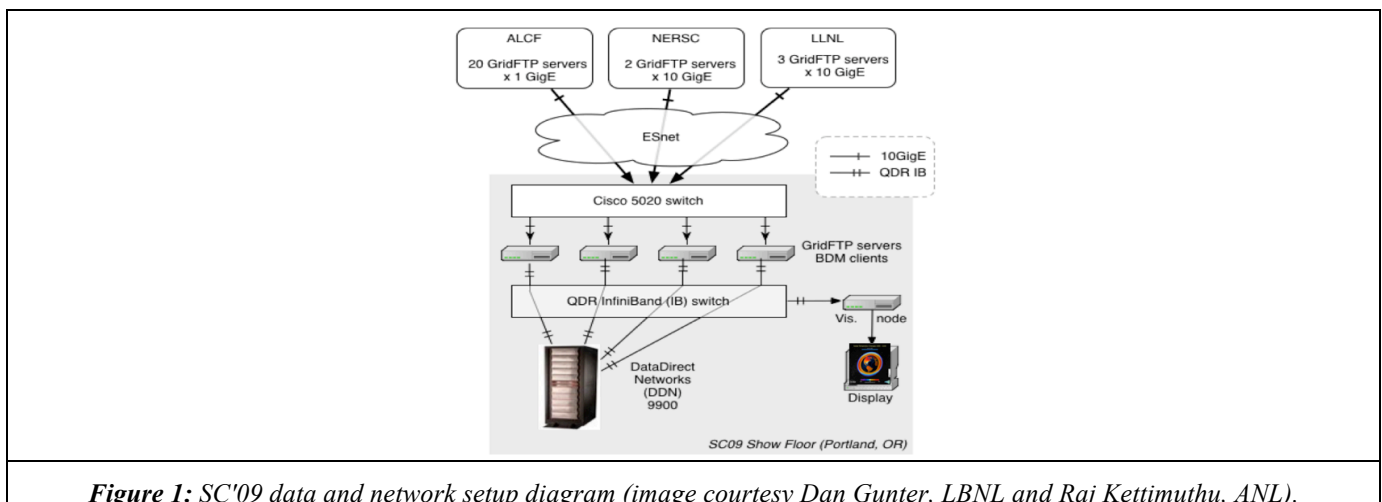
The SC'09 Bandwidth Challenge result was successful. (We took second overall in the competition!). SC'09 effort achieved about 15-Gbps on average and moved over 7 TB of data in an hour from three data sources to the SC'09 show floor. During the effort, a few issues were identified as described in the above sections. It showed our current status in the ESG data replication over the network and what needs to be done in order to make it better for the upcoming CMIP-5 (IPCC AR5) release.

From this effort a paper was generated titled, "Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network". To view this paper, please visit the following URL: http://esg-pcmdi.llnl.gov/publications_and_documents/HPDC10_BWC_Final.pdf/view.

1.1.3 *Data Download Highlights*

In order to access and download data, ESG users are required to have an ESG account at one of the three ESG sites. Many types of data (e.g., CMIP-5, CCSM, POP, PCM, NARCCAP, and C-LAMP) are available for free download at one of the three ESG Gateways located at: PCMDI,

https://esgcet.llnl.gov:8443/index.jsp; NCAR, http://www.earthsystemgrid.org/; and ORNL, https://esg2.ornl.gov:8443/. The graphs below illustrate that use of the ESG-CET continues to grow at a rapid pace as indicated over the last five years. Portals are recording data download volumes in the tens of TBs per month, and the cumulative total since January 2005 has now reached one petabyte (PB) of data downloaded. This certainly is a huge milestone!

Figure 2: *In February, ESG has reached the 1 PB download milestone!*

Figure 3: *The cumulative user-ship of the ESG-CET portals has continued to show growth since 2004. The total number of users by the end of 2010 is very likely to be 20,000. In a given month, between 500 and 800 users are active—up to 30 or more per day.*

Figure 4: *ESG-CET provides support via email (esg-support@earthsystemgrid.org), and tracking the number of e-mails per month shows that the user community is well engaged with ESG-CET. An average day sees between 3 and 4 e-mails, with spikes correlated to releases of additional data sets and codes to the ESG-CET community. Many of these e-mails are related to questions regarding the served data, as well as standard support questions involving resetting usernames and passwords.*

1.1.4 *ESG-CET is Positioning for CMIP-5 (IPCC AR5) Data*

Multiple operational release candidates of the ESG software were scheduled to be deployed and released in March (see *Table 1*) with the goal of replacing the current ESG portals at NCAR (sometime in April) and at PCMDI (later in Spring – early Summer 2010). This software distribution includes major improvements and new functionality in the areas of data versioning, semantic metadata search, data download workflows, database content migration, and metadata exchange.

At the same time, the ESG team was working towards enabling other features to fully support the requirements imposed by the upcoming CMIP-5 data archive, including use of the Data Reference Syntax (DRS) specification, support for data replication, and ingestion and management of detailed model metadata (in collaboration with the Earth System Curator project).

Developers at ANL, British Atmospheric Data Centre (BADC), and NCAR have collaborated on the design and implementation of a new ESG security infrastructure that will replace the token-based system soon after the first ESG software official release. Components of this infrastructure include the Gateway Security Assertion Mark-up Language (SAML)-based Authorization Service, and the DataNode OpenID Relying Party web application and security filter plug-ins for the Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server (TDS).

1.1.5 *LLNL ESG CMIP3 Portal and R&D Highlights*

LLNL has close connections to key climate and science application organizations. The Program for Climate Model Diagnosis and Intercomparison (PCMDI) at LLNL has lead data-management responsibilities for CMIP-5 (IPCC AR5) and other model intercomparison program archives. In addition, the PCMDI team has lead data-management responsibilities for other science applications within the LLNL environment and in private industry. Our team also works closely with other U.S. agencies and institutions to define and implement new technologies, especially those in the areas of data management, distributed computing, networking, and analysis. Note: a particular strength of our team (as well as the entire ESG team) is its close connections to an array of related national and international efforts sponsored by DOE and IPCC as well as the National Science Foundation, National Oceanic and Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), and other organizations in the areas of distributed data, science, and knowledge. These relationships and interactions offer a unique opportunity to leverage major programs across agencies in the support of DOE's research missions.
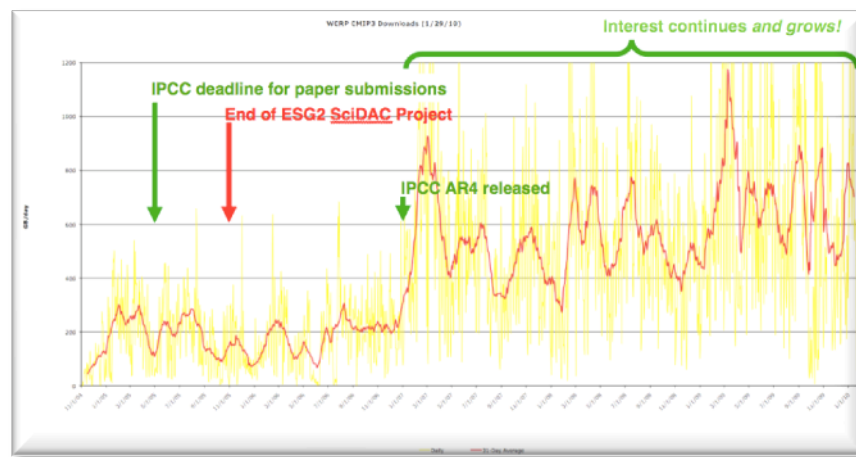
Figure 5: Climate change research is increasingly data-intensive. The picture shows data access from November 2004 through March 2010. Although the IPCC report was released in 2007, interest for the data continues to grow—downloads average over 700 GB per day. Data used by researchers, model developers, policymakers, health officials, etc. for impact and adaptation study and mitigation of climate change are fueling the demand for the data. We suspect similar demands for the forthcoming CMIP-5 data when it is released this summer.

1.1.6 *NCAR ESG Gateway Portal R&D Highlights*

The first operational version of the Gateway software was scheduled to be deployed and released by the end of March, with the goal of replacing the current ESG portals at NCAR (in April) and at PCMDI (later in Spring 2010). This software distribution includes major improvements and new functionality in the areas of data versioning, semantic metadata search, data download workflows, database content migration, and metadata exchange.

At the same time, the ESG team is working toward enabling other features to fully support the requirements imposed by the upcoming CMIP-5 data archive, including use of the DRS specification, support for data replication, and ingestion and management of detailed model metadata (in collaboration with the Earth System Curator project).

Developers at ANL, BADC, and NCAR have collaborated to design and implement a new ESG security infrastructure that will replace the token-based system soon after the first Gateway release. Components of this infrastructure include the Gateway SAML-based Authorization Service, and the Data Node OpenID Relying Party web application and security filter plug-ins for the THREDDS Data Server.

1.1.7 *ORNL ESG Gateway Portal R&D Highlights*

ORNL has deployed a next generation architecture to support ESG-CET. Deployed within the National Center for Computational Sciences, this architecture includes: over 200 TB of high performance disk storage, multiple servers for ESG data and Gateway nodes, a high performance InfiniBand System Area Network, access to our High Performance Storage System (HPSS) archive with over 20 PBs of capacity and 10-Gbps connectivity to the Energy Sciences Network (ESnet). Using this new architecture, ORNL has deployed ESG-CET Data and Gateway node software stacks (Release Candidate 1). In our testbed environment, ORNL is running ESG data and Gateway node software stacks and has published the entire C-LAMP archive, select observational data sets, and other legacy data sets (in excess of 70 TB of data in over 270 thousand files). This work has prepared ORNL for a successful transition to the next generation ESG that will provide the climate science community access to CMIP-5, C-LAMP, and other legacy data sets.

1.1.8 *LANL ESG Node Highlights*

LANL new hardware, Oceans11, was installed on LANL's network with all of the hardware configured for ESG-CET to include 11 TB of usable Direct Attached Storage. Processing and configuration of software has begun and the new installation of the ESG Data Node software stack is being put into place for the 1.0 software release of the ESG enterprise system. To help process Oceans11 model data, a netCDF reader (netCDFPOP) has been developed for ParaView and a version of ParaView with a netCDFPOP reader has been installed on ORNL's Jaguar supercomputer. To this end, the netCDFPOP reader has been rolled into the ParaView source code and scheduled to be release in future minor release versions of ESG. Work on the parallel netCDF reader for ParaView has begun.

1.1.9 *LBNL Berkeley Storage Manager (BeStMan)*

The Berkeley Storage Manager (BeStMan) is a new implementation of the Storage Resource Management (SRM) used in the ESG. BeStMan's architecture is based on a modular design that makes it easy to interface to various storage systems, including Mass Storage Systems (MSSs). Versions of BeStMan have been developed for HPSS at LBNL/NERSC, for HPSS at ORNL, and for MSS at NCAR. These versions of BeStMan deal with different security mechanisms. They have been developed over the last two years, and in the last six months they have been deployed for inclusion in the new ESG federated system. The following tasks were accomplished:

- LBNL/NERSC, NCAR and ORNL BeStMan servers have been installed, passed initial tests, and are now running well.

- They are being continuously tested for underlying HPSS and MSS accesses through new ESG Gateway portal.

- BeStMan servers at all three sites mentioned above have been used by the ESG Metadata Publishing tool for browsing all the file information for files stored on deep storage in order to populate the metadata catalogues at the Gateway. Support for this activity is continuing.

1.1.10   *PMEL Product Delivery Services Highlights*

At the heart of PMEL's contribution to the ESG-CET lies the Live Access Server (LAS), a development activity with origins that date back to 1993. LAS is an XML-configurable workflow engine that has been utilized in support of widely varied climate applications. The introduction of LAS into the ESG in 2006 brought the accomplishments of those efforts into the ESG project and expanded the positive impacts of ESG-specific enhancements to LAS back to those other communities. PMEL has added many enhancements to the model-intercomparison capabilities of LAS and work in this area is on-going. The intercomparison tool (the Visualization Gallery known as vizGal) underwent many behind-the-scenes enhancements that make the client much more efficient and effective. For example, enhancements allow the client to only update the plots displays when a change to the interface affects that panel instead of depending on server-side cache to optimize the refreshing of plot panels. Additionally, vizGal and all of the other LAS components are now benefiting from the completion of the OpenLayers/Google Web Toolkit geographic selection tool and reference map mentioned in the previous report. These ESG-funded enhancements continue to benefit many significant climate and earth science projects that utilize LAS. One of these enhancements will help to pull a wealth of NOAA gridded data (much of it observational in nature) into the ESG/CMIP-5 framework through NOAA's United Access Framework (UAF) project. UAF is a NOAA-wide project that integrates data from across NOAA into a single access framework that builds upon OPeNDAP data access and LAS for data exploration and analysis. Benefits also continue to accrue to other projects using LAS such as the Surface Ocean Carbon Atlas (SOCAT), the Observing System Monitoring Center (OSMC), the Hybrid Coordinate Ocean Model (HYCOM) project and the Global Ocean Data Assimilation Experiment.

1.1.11   *ANL Security, Data, and Services Highlights*

ANL has worked with the ESG team to provide single sign-on access across the ESG federated sites, and demonstrated interoperability between multiple sites. OpenID and MyProxy server were used to achieve single sign-on across the web and Public Key Infrastructure (PKI) realm. This feature will be released as part of the ESG-CET 1.0 production release in April 2010.

ANL has also led the design of the security architecture for the LAS to ensure that data download access policy is enforced for data product generation and visualization using the server. This work, done in collaboration with the Gateway and Product Services team, leverages the current ESG authentication and authorization infrastructure to secure access to the LAS server. The solution will be deployed as a part of the next ESG milestone release.

In addition to user download services, ANL designed and implemented an attribute-based authorization mechanism for bulk-data transfer to replicate core climate data sets. The re-designed mechanism reduces administrator burden and usability by minimizing the provisioning and configuration requirements. The new mechanism authorizes access based on a user attribute agreed across the federation, rather than requiring configuration of the identities used for replication. With this option, any administrator can be assigned the attribute by a central site without having to propagate the changes to the administrator's identity to all ESG sites.

1.1.12 *USC/ISI Mirroring Highlights*

The data replication team at ISI has made extensive progress on the design of the replication client for ESG. We held a series of teleconferences in Fall 2009 and Winter 2010 to better understand replication requirements and to identify missing pieces of functionality, which we have been working to develop.

Our progress has included development of a draft application programmer's interface (API) for the replication client, which was then refined and modified during extensive discussions with both the ESG team and our collaborators in the Global Organization for Earth System Science Portals (GO-ESSP) team.

We also held a series of design meetings to understand the impact of replication and mirroring of ESG data sets on the design of gateways and data nodes. Based on these discussions, the NCAR team developed a scheme for sharing metadata about replicated data sets among the gateways so that replicas could be discovered and accessed.

1.1.13 *Rensselaer Polytechnic Institute*

We have begun preliminary work testing the ESG-CET authorization client library with intent to integrate it into netCDF-based clients. At present, the library only works under the Linux operating system, and thus the GridFTP module will only work under Linux as well. This restriction will need to be overcome as it hampers the benefit of putting the authorization in the client, it must now happen in the ESG infrastructure.

We have substantially contributed to the OPeNDAP Hyrax 1.6 beta source release with support for NcML, which is a key requirement for aggregation support for ESG-CET. We are in the process of installing the THREDDS Data Server (TDS) for baseline functionality comparison between TDS and Hyrax-g for the THREDDS catalogs.

We are continuing to develop the build and test installation software for the ESG-CET customizations for Hyrax-g as well as pay close attention to code security via static code analysis and code audits.

2   Overall Progress

During this reporting period, progress was made in key areas that are necessary to meet ESG-CET objectives, goals, and milestones. This section provides greater technical depth and presentation of the components needed for the beta release.

Gateway Software Development

In the past six months, the development of the Gateway software stack has been guided by two main goals: the imminent transition of the two ESG operational portals at NCAR and PCMDI to the new infrastructure (scheduled for Spring 2010), and support for the upcoming CMIP-5 model output streams (anticipated for Summer 2010). The Gateway functionality has been augmented and revised in many respects, including the following major areas of development:

- *Faceted Search.* A completely new user interface for the ESG-faceted search has been developed, and featured prominently on the home pages of the PCMDI and NCAR Gateways. The new interface follows paradigms that are becoming common across many e-business web sites, thus facilitating the user experience as well as increasing the scalability and maintainability of the application.

- ***Data Reference Syntax (DRS).*** The Gateway application has been enabled with the capability to ingest and expose metadata conforming to the DRS specification that will be used to organize and index the CMIP-5 data archive. This includes parsing of the DRS information from the THREDDS catalogs, persistence in the relational database, and configuration of DRS facets in the metadata query interface.
- ***Data Versioning***. The Gateway underlying domain model has been heavily refactored to support data versioning (i.e., the need to publish, update and retract new versions of the data in response to quality control results). Versioning support had implications at all levels of the Gateway software: the relational object model, the publishing services, the metadata query services, and the user interface.
- ***Data Migration***. A procedure for migrating the relational database content when upgrading the underlying schema has been setup based on the open source Liquibase project. This procedure addresses a critical operational need since it will allow Gateway administrators to upgrade to a new major software release without having to republish all the data or ask users to register again.
- ***Authorization Service.*** In collaboration with ANL, the Gateway application has been augmented with an embedded Authorization Service, which can deliver Security Assertion Markup Language (SAML) signed assertion in response to authorization queries by remote clients (in particular, the DataNode plug-ins for the GridFTP and TDS servers).
- ***Metadata Exchange.*** The Gateway infrastructure for metadata exchange, based on the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH) protocol, has been updated to support versioning and replication, as well as the capability to execute selected harvesting by project so that a Gateway needs only import those records that are relevant to its user community, and not others.
- ***Data Download.*** The user workflows for selecting and downloading the data have been thoroughly revised to make them more friendly, efficient, and to support a variety of options including selection across data collections and gateways, retrieval of files from deep storage via BeStMan, generation of wget scripts and integration with the DML desktop client.
- ***Model Metadata.*** The Earth System Curator project has continued to work together with ESG to expand the Gateway functionality for ingesting and servicing model metadata, including the capability to connect data sets to the models and simulations that produced them, full handling of the CMIP-5 conformance properties, ingestion of CIM metadata from the METAFOR Questionnaire application, and several improvements to the user interface (the model "trackback" pages). Progress in this area was guided by a series of demonstrations that the Curator project organized to solicit continuous feedback from national and international partners, such as the GO-ESSP and METAFOR communities.
- ***Gateway User Interfaces.*** The Gateway User Interface has been revised and improved throughout the site, including major changes in the pages for metadata search, data download, model trackback, user and group administration.

Gateway Software Distribution and Management

The Gateway software has been released, deployed, and tested according to a carefully planned schedule aimed at progressively enhancing the functionality of the application, while at the same time soliciting feedback from a progressively larger user base. Starting in October 2009, the Gateway team released several alpha, beta, and release candidate versions with the first production release 1.0 expected by the end of April 2010. The main goal was to replace the current operational ESG portals at PCMDI and NCAR. Additionally, other releases are planned for the upcoming months with the various goals of

enabling federation-level services (1.1), upgrading the overall security infrastructure (1.2), enabling LAS visualization services (1.3), and fully supporting the CMIP-5 requirements (1.4). Currently, the Gateway software is installed and operational at three sites: NCAR, PCMDI and ORNL. An earlier version has been installed for evaluation by the BADC collaborators.

The Gateway team has adopted the Atlassian software engineering suite, employing Jira for planning software releases, tracking features and bugs, and providing development provenance and automatic release reports. We are now well into the phase of utilizing a continuous integration environment, with testing for each set of commits, and reports from a nightly build via the Atlassian Bamboo component. This system has been recently opened to outside collaborators and will be used to plan and prioritize work for all future releases. We have also invested up-front time in developing a database migration capability, which serves a critical rapid-release-cycle function, along with the beginnings of a test harness that provides automated functions for exercising the Gateway, and assessing the viability of new releases. During the next reporting period, the Gateway software will be released as an open source initial (OSI)—a qualified open source product along with a well-defined community process for further development and enhancement.

Data Node Software Development

In the ESG Data Node software stack (see **Figure 5**), we have begun the development of a coordinating entity called the *Data Node Manager*. The architecture of this component is unique from the other components because it addresses cross cutting concerns that span all the constituent elements of the stack. The *Data Node Manager* logs, monitors, and collects metrics over the entire software stack. It has been designed from conception to be highly fault tolerant and perform under load with graceful degradation properties.

Figure 6: *This diagram shows the beginning operations performed by the Data Node within the ESG infrastructure.*
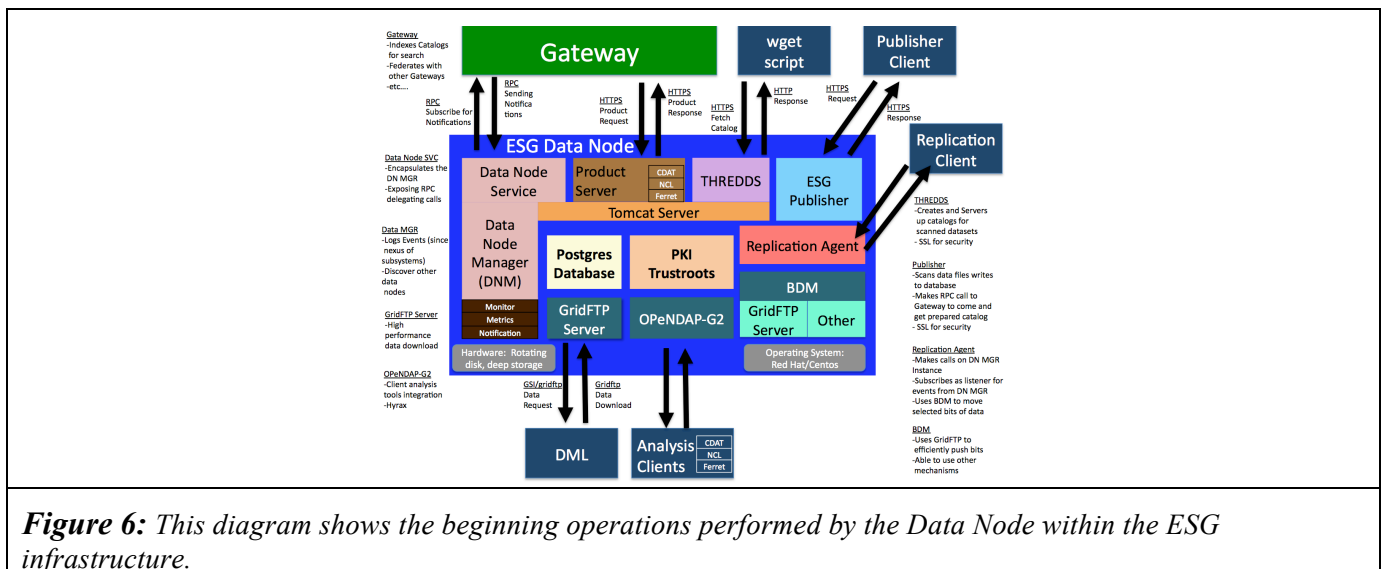
In the ESG infrastructure, the actual data holdings reside on a potentially large number of federated nodes collectively referred to as the ESG Data Node. In addition to hosting those data, the Data Node includes the metadata services needed to publish data to portals and execute data-product requests through these portals. Personnel can set up nodes at local institutions, and the single ESG Gateway

serves data requests to many associated nodes. For example, more than 30 institutions are expected to operate data nodes as part of the IPCC Coupled Model Intercomparison Project, phase 5 (CMIP-5).

The Web application for the ESG Gateway is instrumented with the machinery necessary to parse THREDDS data catalogs, create hierarchies of domain model objects, and persist them to the relational database. This functionality is available as a Hessian Web service that can be invoked by the publishing application running on the Data Node, requiring secure mutual authentication via digital certificates. The ESG infrastructure includes the following components, which are also shown in *Figure 5*:

- *Support for high-performance data download client.* The Gateway application is augmented with the Data Mover-Lite (DML) client, which supports high-performance, multiple-file download. Users can download and configure DML directly through the Gateway and then start it via a file list in XML format. With this system, users can request access to files stored on a local rotating disk or those in deep storage archive, which are then transferred to the Gateway cache.

- *Data replication.* Data replication is an important facet of international collaborations such as ESG-CET. No one facility can host the accumulated holdings for the data-intensive climate research community; therefore, archives at one institution are replicated at a collaborator's site via software links. ESG has a robust data replication client, developed and improved by analyzing replication use cases and requirements to ensure that the IPCC archive stored at LLNL is available to all of the participating the European data centers.

- *Semantic search.* An alternative user interface resides on top of the ESG semantic services for data search and discovery. This interface more closely resembles the general characteristics of standard business Web sites and thus may be a more intuitive, easier-to-use tool for users.

- *Modeling metadata.* The ESG collaboration is supporting the Earth System Curator (ESC) project by developing the full infrastructure so that it accurately captures and displays model metadata within the Gateway Web application.

- *Federated authentication.* ESG is designed as a federated system allowing user access via the ESG Gateway and supporting interoperability with other non-ESG partner data centers. The ESG infrastructure includes the OpenID protocol, which supports cross-site authentication between the many gateways as well as with European Web portals. OpenID is a single-sign-on technology that allows users to register at only one site, where their credentials are stored, and then carry their authenticated identity as they navigate and access data throughout the ESG federation.

- *User attribute services.* ESG, in collaboration with its European partners, chose the SAML as the enabling technology to exchange user attributes and access control information among sites. Each ESG gateway or partner data center will deploy a SAML-based attribute service, which other gateways can securely query to retrieve attribute information about a specific user. This information is required to authorize users both to access specific data sets controlled by a group at another site and to store complete data access metrics.

- *Data publishing operations.* After the Data Node software stack is installed, users can run software scripts required to publish all current ESG data holdings (i.e., climate model simulations and observations) to the Gateway. This support for full publication ensures that users have access via the Data Node software stack application to all holdings, including nonstandard model runs, multiple deep-storage archives, and multiple data access services.

The ESG Data Node software stack is distributed nationally and internationally via a virtual machine (VM), a software platform that allows a complete and sovereign operating system (a guest OS) to run as an application inside another operating system (the host OS). The guest OS executes software applications identically to a physical machine. The VMs contain the fully installed CentOS (with requisite libraries and other functionality installed) and the Data Node software stack. (This setup eliminates the need for the system to check all software prerequisites). Hardware virtualization via VM improves security, insulates the hardware from attack and user error, offers code portability and ease of backup, and protects the system against potential software conflicts. The following links to the CentOS ESG-CET Data Node VM installation files describe VM configuration in more detail:

- http://rainbow.llnl.gov/dist/datanode/ESG_CentOS_Linux_2.6.x_kernel.v4.dn.tgz
- http://rainbow.llnl.gov/dist/datanode/ESG_CentOS.v4.dn.md5

Data Node Security Infrastructure

ANL, NCAR, and BADC have closely collaborated to define and implement a new ESG security architecture that would support secure data access by browsers and rich analysis and visualization clients throughout the federation. The new architecture is based on technologies such as OpenID, SAML, and X509 certificates and is scheduled to replace the current ESG token-based access in Spring 2010. As part of this collaboration-wide effort, NCAR has delivered the following components:

- A Java-based OpenID Relying Party (ORP), which is a web application (running within a Tomcat servlet container) that is responsible for either validating a user certificate or redirecting a user browser to an OpenID Identity Provider, with the net effect of establishing a secure authentication cookie that can be used by other DataNode software components as a proof of the user's identity.

A servlet filters infrastructure that can be used in conjunction with the ORP application to secure a generic Java-based data server, such as the THREDDS Data Server that will be part of the standard ESG DataNode application stack. Existing implemented components include an authentication filter to consume the cookie set by the ORP and an authorization filter that contacts the Gateway Authorization Service.

Cyber Security

Enhancements to the security architecture of data transfer services, such as GridFTP server and TDS, have been completed to ensure compliance with standards and interoperability across the different federation sites. The ESG data servers used token-based authentication and authorization, which other than being non-standard allows any user with the token to access the data. ANL worked on designing a security solution that replaced a non-standard and less secure mechanism of authorizing access to the data files. We worked with the ESG Gateway team to provide a SAML-based Authorization and Attribute service. These services use the standard constructs to communicate authorization decisions and user attributes. These are being integrated into the security framework of TDS and GridFTP server so that the data transfer services callout to these services for a decision about whether a particular data set can be downloaded by a user. Furthermore, we worked on a solution to optimize the GridFTP authorization callouts so that multiple files can be authorized with a single callout.

DataMover-Lite (DML)

Changes in requirements for DML came up during this period for its operation with the authorization plug-in in GridFTP servers. The main requirement is to pass full URL as a site command so that the

authorization plug-in module can find the port number for the files requested. DML version 3 has been developed, integrated with the new ESG authorization model and plug-ins, and tested extensively with GridFTP servers. Both the stand-alone version and the Web-start version have been developed and tested.

Bulk Data Mover (BDM)

The BDM has been developed and the first working version was released in November 2009. Currently, it is undergoing extensive testing and enhancements.

The BDM is responsible for the successful replication of large data sets. Climate data sets are characterized by large numbers of small files. To handle this issue the ESG uses the BDM software as a higher-level data transfer management component to manage the file transfers with optimized transfer queue and concurrency management algorithms.

The BDM can accept a request composed of multiple files or an entire directory. The files or directory are described as Universal Resource Locators (URLs) that indicate the source sites that contain the files. The request also contains the target site and directory where the replicated files will reside after successful transfers. If a directory is provided at the source, then the BDM will replicate the structure of the source directory at the target site. The BDM is capable of transferring multiple files concurrently as well as using parallel transmission control protocol (TCP) streams. The optimal level of concurrency or parallel streams is dependent on the bandwidth capacity of the storage systems at both ends of the transfer as well as achievable bandwidth on the wide-area-network (WAN). Setting up the level of concurrency correctly is an important issue, especially in climate data sets, because of the small files. Concurrency that is too high becomes ineffective (high overheads and increased congestion), and concurrency that is too low will not take advantage of available bandwidth. A similar phenomenon was observed when setting up the level of parallel streams in a single file transfer, such as GridFTP.

The BDM is designed to work in a "pull mode", where the BDM runs as a client at the target site. This choice is made because of practical security aspects: site managers usually prefer to be in charge of pulling data rather than having data pushed at them. However, the BDM could also be designed to operate in a "push mode" or as an independent third-party service. Because a large scale data replication can take a long time (from many minutes to hours and even days) the BDM must be an asynchronous service. That means that when a replication request is launched, a "request token" is returned to the client. The client should be able to use that request token to check the status of the request execution at any time. Another obvious implication to the long lasting nature of large scale replication is the need for automatic monitoring and recovery from any transient failures, which is an important part of the BDM's design.

Multi-phase Transfer Request Management

The tasks that the BDM performs to accomplish a successful replication are organized into three phases, as shown in *Figure 7*. The initialization phase plans and prepares file replications from the data source to the local target storage. It includes the following tasks: 1) Storage allocation verification at the target site. This requires collecting the total data size from the source site. 2) Generating a request plan. The plan includes the initial level of concurrency, number of parallel streams, and buffer size for the request. 3) Returning an initial request estimation to the client. 4) Mirroring the directory structure of the source at the target site. It then generates an execution plan that includes pair-wise source-to-target URLs for all the files to be replicated. This is used by the execution phase.
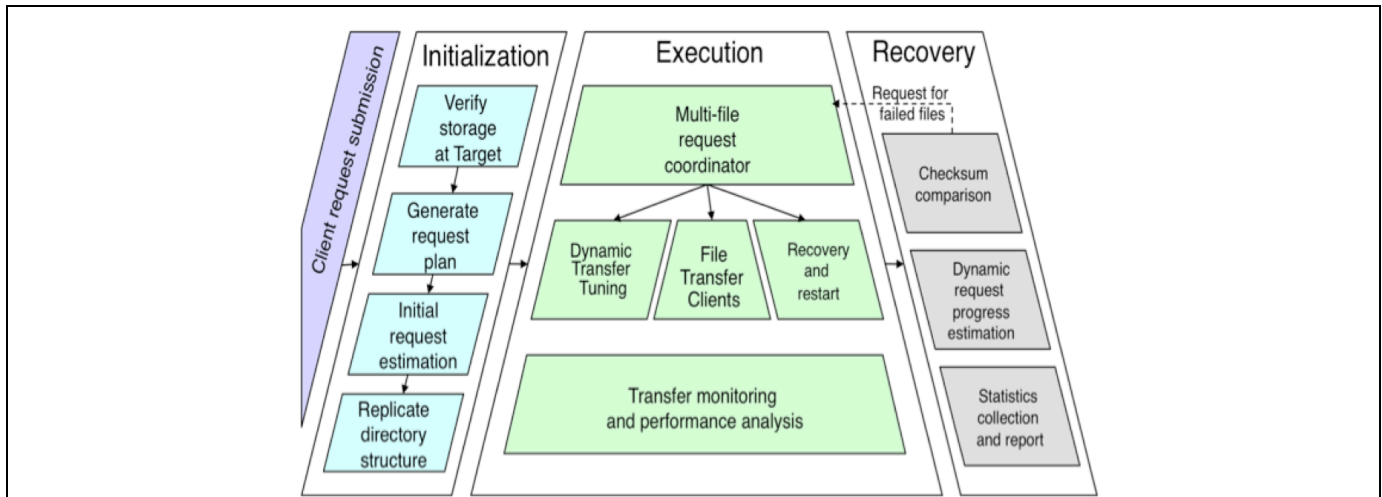
Figure 7: *The design of the BDM.*

The Execution phase transfers the requested files, while monitoring and analyzing transfer performance for dynamic adjustment on the transfer properties. It consists of four modules. 1) The Multi-File Request Coordinator uses the information from the "execution plan" and transfer properties including the concurrency level, and accordingly instantiates the file transfer client. 2) The File Transfer Client can support any transfer protocols or services preferred by the virtual organization. Supported transfer protocol could be GridFTP, HTTPS, secure copy (SCP), secure FTP (SFTP), etc. 3) The Recovery and Restart module continuously monitors the health of the system and the files being transferred. If a transient error occurs it waits for the system to recover and depending on the transfer protocol used, either removes the partial files transferred and reschedules the transfer or continues the transfers from the point of interruption. For example, GridFTP allows partial transfers to be resumed. 4) The module responsible for monitoring and adjusting concurrency collects dynamic transfer performance, and if significant discrepancies from the estimated performance are noticed, it adjusts the number of concurrency and parallel streams.

The Recovery phase interacts dynamically with the components of the execution phase to validate the completed request by collecting statistics, generating dynamic progress estimation on-demand, and validating transferred files at the end of the request. It has three functions. 1) It collects statistics from the execution of the replication request. 2) It generates dynamic progress estimation on-demand when a client asks for request progress status. This module needs the information on file transfers that completed, are in-progress, or are pending, as well as bandwidth usage statistics and estimation. 3) The file validation module can be running as soon as files are transferred, or at the end of the request, depending on the site preference. The reason for preferring file validation by checksum comparison after all transfers complete is that calculating checksums is computationally intensive and may disturb the running transfers. This module is also responsible for re-submitting files whose checksums indicated data corruption.

Transfer Queue Management and Balanced Concurrency

The BDM achieves high performance using a variety of techniques, including multi-threaded concurrent transfer connection management, transfer queue management and single control channel management for multiple data transfers, while the GridFTP library supports data channel caching and pipelining.
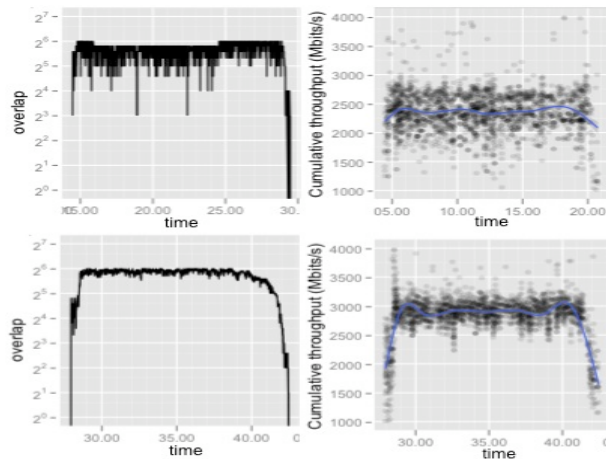
Transfer queue management and concurrency management contribute to higher transfer throughput, including both on the network and from the storage system. When there are many small files in the data set, continuous data flow from the storage into the network can be achieved by pre-fetching data from storage onto the transfer queue of each concurrent transfer connection. This overlapping of storage Input/Output (I/O) with the network I/O helps improve the performance.
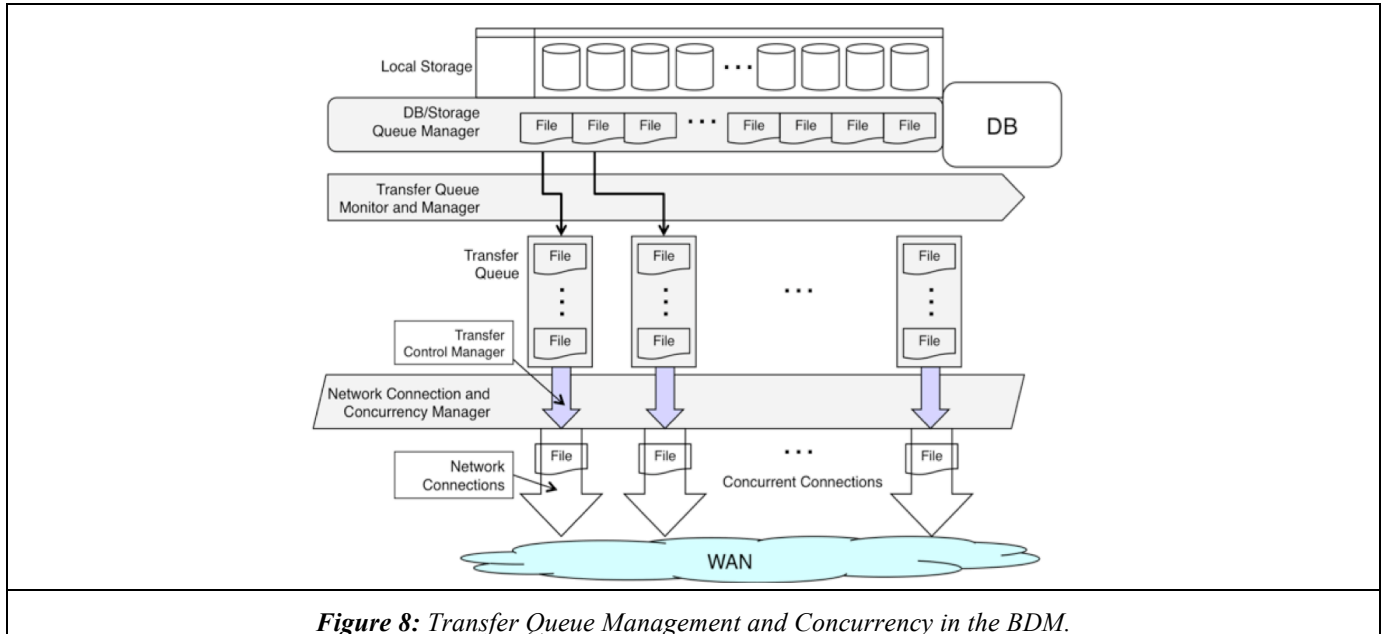
Figure 8: *Transfer Queue Management and Concurrency in the BDM.*

As shown in *Figure 8*, BDM manages a database (DB) queue for balanced access to DB from the concurrent transfer connections, and also manages the transfer queues for concurrent file transfers. Each transfer queue checks a configurable threshold for the queued total files size and gets more files to transfer from the DB queue when the queued total files size goes below the configured threshold. Default threshold is set to 200 megabytes (MB) based on the file size distributions in data sets. *Figure 9* shows the results from the effect of transfer queue and concurrency management. When the transfer queue and concurrency are well managed (shown in the bottom left plot of the figure), the number of concurrent data transfers is consistent over time, compared to the ill- or non-managed data transfers (shown in the top left plot of the figure), and it contributes to a higher overall throughput performance (shown in the bottom right of the figure).

For further optimization in the transfer queue management, the order of the transfer files based on file sizes is considered in concurrent transfer connections. Also, the concurrent transfers are balanced across multiple source transfer servers, when multiple transfer servers are available at the data source. In the future, network delay will be considered in the calculation of concurrency and parallelism in BDM file transfers.

Figure 9: Results from transfer queue and concurrency management in BDM, showing ill- or non-managed queue and concurrency on the top row and optimized queue and concurrency on the bottom row, for the number of overlapping concurrent transfers on the left and the transfer throughput over time on the right.

Data Replication

The ISI team is working with the LLNL team on data publication issues for replicated data sets, including identifying necessary functionality that had to be added to the publication client. The LLNL team responded by modifying the publication client to add metadata to identify published data sets as replicas. They also developed an API for use in data replication that includes the ability to separately scan a mirrored directory, create a THREDDS catalog for it, and publish the mirrored data set to a designated gateway node.

The ISI and LLNL teams have also worked together on operational issues related to data replication. These include the requirement to run a BDM-enabled GridFTP server at LLNL for us to download the data in our testing. We identified issues related to this GridFTP server, which needs to run on a dedicated machine because of the resources it consumes. We also agreed on a security model for this GridFTP server since it will use the role-based authorization being developed by the security team rather than the older token-based authorization scheme.

The ISI team has also done extensive testing of the replication client and the publication client developed by LLNL. ISI staff have tested the documentation and procedures for installation of a Data Node, both manually using the instructions provided by the LLNL team and using the recently developed installation scripts. The ISI team has worked to identify problems and bugs in these instructions and scripts and relayed this information to the LLNL team so that it can improve the documentation and scripts.

Finally, the ISI team downloaded data from the ESG Web portal to the local Data Node and published that data to the PCMDI Gateway using the ESG publication client. These newly published data sets were then discovered and downloaded.

In the next 6 months, the ISI team plans to complete development of the initial replication client functionality and release it to the ESG and GO-ESSP communities.

Product Services

The ESG-CET is intended to serve information products to users representing a broad spectrum of sophistication—from numerical modelers who want access to "raw" model output files and verbatim subsets of model output to climate impacts investigators who want rapid access to these data without the complexities of model-specific coordinate systems and to users who only want to quickly visualize the overall behaviors of models. Here in bullet form are the highlights of the enhancements to LAS made through ESG support:
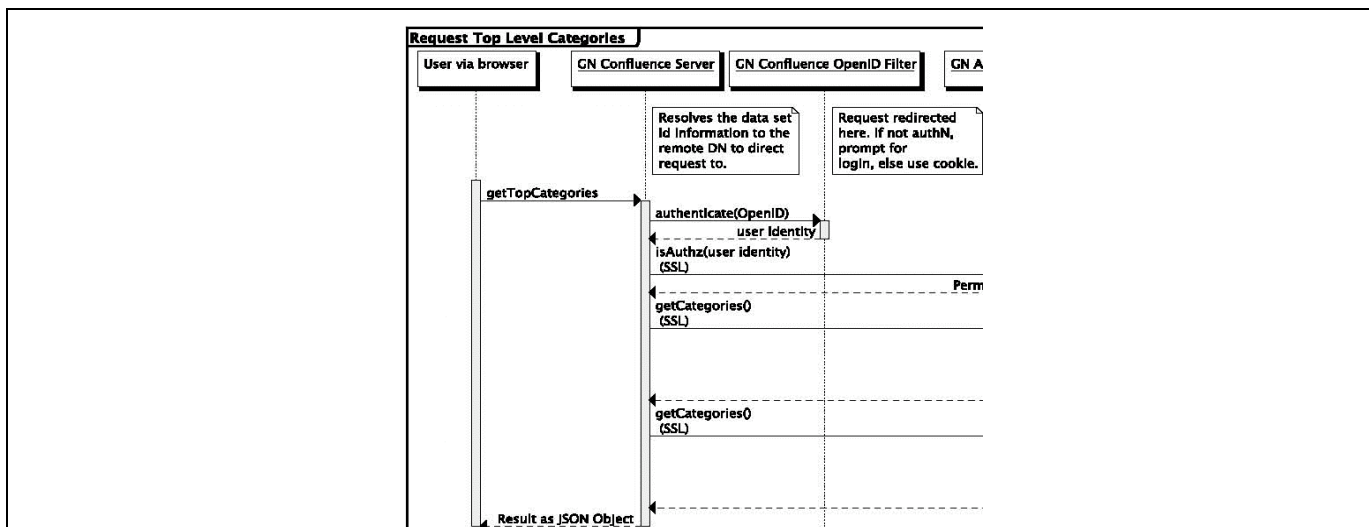
Figure 10: *Section of an authentication and authorization sequence diagram for LAS (full documents available on-line).*

- Integration of LAS into the authorization and access control framework used by ESG-CET (on-going development).

- Data inter-comparison capabilities through the vizGal interface. This interface is under active development to enhance or add capabilities for intercomparison along slices in any dimension, including comparison of time-series, Hoffmuller plots, and vertical section plots.
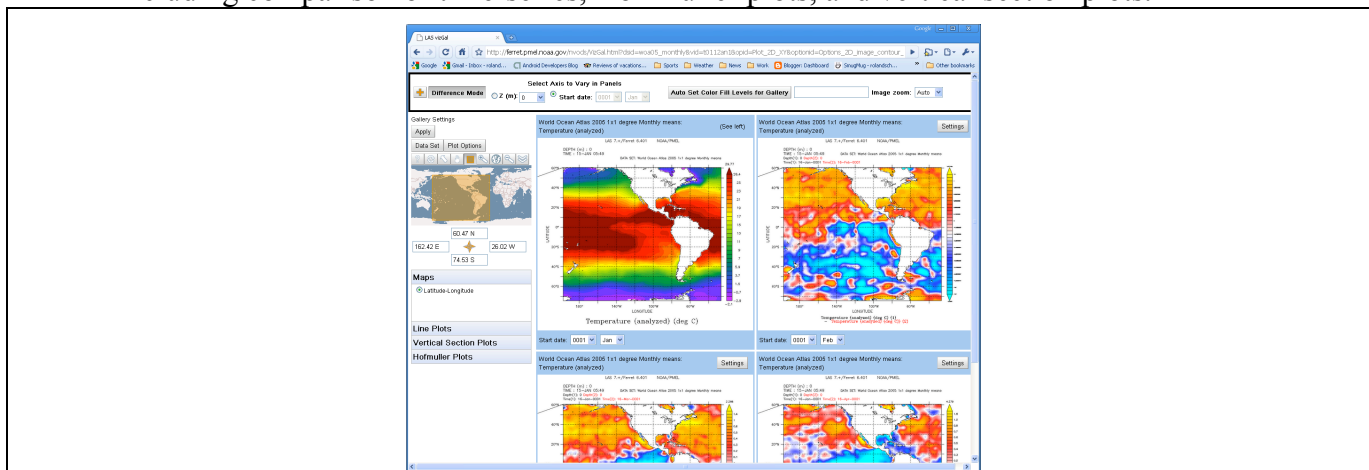
Figure 11: *Bowser displaying the new vizGal interface with multiple views showing model intercomparison.*

- THREDDS catalog "cleaning tools". An important part of the ESG-CET science mission is to make meaningful inter-comparisons between disparate data collections. Unfortunately, many interesting data sources to which models can be compared (i.e., satellite data collections, operational model outputs, and gridded oberservational assimilations) are not kept in well-organized archives with fully developed file-level metadata. To help organize such data collections, we have built THREDDS catalog scanning tools that can identify data sets with excellent metadata from large collections, automatically create aggregations from collections of individual time-series files and create catalogs that reflect only the best of large often jumbled data repositories. These clean catalogs can let scientists from ESG-CET and elsewhere quickly identify data sets, which can be immediately used to further their scientific goals.

3 ESG-CET Group Meetings

The ESG-CET executive committee holds weekly conference calls on Mondays at 9:00 a.m. pacific time. At these meetings, priorities and issues are discussed that make up the agenda for the weekly project meetings held on Thursdays at 12:00 p.m. pacific time via the Access Grid (AG). At the AG meetings, the entire team discusses project goals, design and development issues, technologies, timelines, and milestones. Given the need for more in-depth conversation and examination of work requirements, the following face-to-face meetings will be held during this next reporting period.

4 Collaborations

To effectively build an infrastructure capable of dealing with petascale data management and analysis, we established connections with other DOE Office of Science SciDAC-funded projects and programs at various meetings and workshops, such as the SciDAC Outreach Center Workshop held in San Francisco, California. In particular, collaborations were established with the following groups: TeraGrid Science Gateways, Earth System Curator, NOAA's Global Interoperability Program (GIP), METAFOR, World Meteorological Organization (WMO) - WMO Information System (WIS), Scientific Computing and Imaging (SCI) Institute at the University of Utah, SciDAC VACET, SciDAC SDM, SciDAC CEDPS, Southern California Earthquake Center (SCEC), Tech-X Corp., NASA JPL, NASA Goddard, GO-ESSP, and many others.

2009 Global Organization for Earth System Science Portal (GO-ESSP) Workshop

Dean N. Williams, Steve Hankin, and Don Middleton are three of seven GO-ESSP steering committee members who coordinated the eighth annual GO-ESSP workshop held October 6 – 8 at the Institute for Pharmacy in Hamburg, Germany. In addition, Steve, Don, and Dean chaired workshop sessions.

The GO-ESSP workshop focuses on facilitating the organization and implementation of an infrastructure for full data sharing among a consortium spanning continents, countries, and intergovernmental agencies. This GO-ESSP consortium envisions an environment that allows users open access to PB of model-generated, satellite, and in-situ data including physical, biogeochemical and ecosystem content. All initial ESG Federation partners were present (i.e., LLNL, NCAR, GFDL, BADC, German Climate Computing Centre (DKRZ) and the University of Tokyo). The workshop, in part, covered data security, versioning, and replication concerns and addressed issues of collaboration. By 2011, this organization envisions allowing users open access to PB of multi-model generated data, as well as in-situ, satellite, biogeochemistry, and ecosystems data.

NOAA Global Interoperability Program (GIP)

Dean N. Williams, Don Middleton, Steve Hankin, and Luca Cinquini attended the NOAA-funded GIP kickoff meeting held at the Geophysical Fluid Dynamics Laboratory in Princeton, NJ on November 5–6, 2009. The GIP program promotes coordination of software infrastructure development across agencies, the weather and climate communities, modeling and data services, and research and operational centers.

The Curator project is now hosted under the NOAA GIP program, and it continued a highly productive collaboration with ESG-CET. Curator continued to serve as an active liaison between the EU-based METAFOR project, which is developing a Common Information Model (CIM) for CMIP5/IPCC, and ESG-CET, which will be delivering data to the global community. We have jointly demonstrated new "model trackback" capabilities, which effectively combine the METAFOR questionnaire with the query and browse functions of ESG.

Hybrid Coordinate Ocean Model (HyCOM) consortium (NOAA, Navy, et. al.)

NOAA/PMEL (Steve Hankin, ESG co-PI) is a partner in the Hybrid Coordinate Ocean Model (HyCOM) consortium [http://hycom.rsmas.miami.edu/]. The HyCOM Consortium has developed a high resolution (1/12 degree) operational, global ocean modeling capability under cooperative U.S. Navy and NOAA funding. The HyCOM model presents unique technical challenges through the complicated coordinate system that it employs and its large data volumes, but the needs of HYCOM overlap in many respects with the ocean components of the climate models to be utilized in IPCC AR5. There is a significant and productive two-way technology transfer of technical capabilities developed in support of ESG and technical capabilities developed in support of HyCOM.

NOAA Geophysical Fluid Dynamics Laboratory

The NOAA GFDL Fluid Dynamics Laboratory is an active contributor to AR5 and an active participant in the ESG SciDAC. V. Balaji (Head, GFDL Modeling Systems Group) is a frequent participant and active contributor in ESG telephone conferences and meetings leading to a vigorous bi-directional exchange of ideas and technology. NOAA/PMEL (Steve Hankin, ESG site-PI) shares an Memorandum of Understanding (MOU) with GFDL for the development of the LAS and analysis tools, also leading to an active two-way technology transfer between NOAA and ESG.

NOAA Office of Climate Observations (OCO)

PMEL is the developer of the OSMC on behalf of NOAA/OCO and manages interactive access to the international SOCAT for quality control analysis. Through the PMEL membership in the ESG SciDAC a number of useful collaborative benefits are being explored and are likely to be realized in time for IPCC AR5 work. OSMC and SOCAT are both sources of integrated ocean-climate observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be helping to bring these collections of observations into the ESG framework for the benefit of IPCC scientists and others.

Unidata and the Climate and Forecast Conventions (CF)

Several ESG members play key roles in the development of the CF conventions—the emerging standard for climate model outputs stored in netCDF. ESG is forging a strong collaborative relationship with Unidata, the development organization for netCDF.

US Integrated Ocean Observing System (IOOS)

PMEL is a member of the US Integrated Ocean Observing System (IOOS) Integrated Products Team (IPT). IOOS is a potential source of integrated ocean observations that are potentially useful to IPCC scientists in the evaluation of climate model outputs. PMEL will be collaborating with IOOS to locate climate-relevant U.S. coastal observations and bring them into the IPCC framework.

Global Earth Observation Integrated Data Environment (GEO-IDE)

The PMEL TMAP group put forward at the GEO-IDE meeting that it would lead a small community in the creation of a distributed THREDDS catalog of NOAA gridded data sets. We have already populated this publicly available catalog with data sets from across several different NOAA line offices, including Oceanic and Atmospheric Research (OAR), National Marine Fisheries Service (NMFS), and National Environmental Satellite, Data, and Information Service (NESDIS). This collection provides a rich set of data for inter-comparison and verification with the main ESG-CET collections.

ESG-CET Collaborates with CDIAC and ARM Data Centers at ORNL

The Carbon Dioxide Information Analysis Center (CDIAC) and the Atmospheric Radiation Measurement Program and the ESG-CET team at ORNL have begun a collaborative pilot project to evaluate the feasibility of integrating high-value observational data sets into ESG. This work has resulted in the successful publication of a number of observational data sets, identification of architectural enhancements within ESG to support observational data sets, and the identification of metadata enhancements that will be required for data discovery.

NCDC visit to ORNL to discuss ESG-CET and Observational Data

Thomas Karl, Scott Hausman, John Bates, Eileen Shea, and Russell Vose (National Climatic Data Center) visited ORNL to discuss multiple areas of collaboration. John Bates was particularly interested in ESG-CET to enable derivative data products that fuse observational data sets with climate model data. Further collaborations with Helen Frederick and others at NCDC are ongoing with the goal of providing seamless access to data sets at NCDC and ORNL.

WMO Information System (WIS)

The WMO WIS Program is developing a next-generation globally federated data system that is intended to serve all of the WMO areas (e.g. climate, weather, hydrology, etc.). Don Middleton contributes to several WIS committees and regularly provides briefings that encompass ESG-CET and our progress towards CMIP5/IPCC objectives and climate research in general. Our Gateway technology will provide federation with WIS while also serving as a WIS validation platform.

NSF TeraGrid Science Gateways Program

ESG-CET, NOAA GIP, and Purdue University staff collaborated to propose the development of a prototype Environmental Science Gateway, and was granted support from TeraGrid's Science Gateway program for two years. The basic objectives of this project include developing a Gateway-based model execution capability, a supporting workflow system that can operate on the TeraGrid, and data publishing and harvesting capabilities that can interoperate with an iRODS-based archive as well as disk-hosted model output.

5   Outreach, Papers, Presentations and Posters

Outreach activities, papers, talks, and posters presented during this time period:

Outreaches Activities

Papers:

5.1.1   *SC-09 High Performance GridFTP Transport of ESG Data*

The Supercomputing 2009 (SC'09) Bandwidth Challenge entry titled, "High Performance GridFTP Transport of Earth System Grid (ESG) Data", demonstrated high performance GridFTP transport of climate data from multiple Department of Energy laboratory locations to the targeted SC showroom floor. The transferred multi-terabyte data consisted of a small portion of the multi-model Coupled Model Intercomparison Project, Phase 3 (CMIP-3) data set used in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4). SC'09 conference proceedings.

5.1.2   *Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network*

Raj Kettimuthu, *Alex Sim*, Dan Gunter, Bill Allcock, Peer-Timo Bremer, John Bresnahan, Andrew Cherry, Lisa Childers, Eli Dart, *Ian Foster*, Kevin Harms, Jason Hick, Jason Lee, Michael Link, Jeff Long, Keith Miller, Vijaya Natarajan, Valerio Pascucci, Ken Raffenetti, David, Ressman, *Dean Williams*, Loren Wilson, and Linda Winkler, "Lessons learned from moving Earth System Grid data sets over a 20 Gbps wide-area network", http://esg-pcmdi.llnl.gov/publications_and_documents/HPDC10_BWC_Final.pdf/view.

5.1.3   *Experience Report on the Supercomputing 2009 Bandwidth Challenge from the Perspective of the Earth System Grid*

Alex Sim and Dean N. Williams, "Experience Report on the Supercomputing 2009 Bandwidth Challenge from the Perspective of the Earth System Grid", Supercomputing Conference '09 Bandwidth Challenge, December 2009, http://esg-pcmdi.llnl.gov/publications_and_documents/ExpReports-ESG-SC09_final.pdf/view.

Talks:

5.1.4   *DOE BER Climate Change Modeling Program Science Team Meeting*

Dean N. Williams presented, "The Earth System Grid Center For Enabling Technologies (ESG-CET): Serving Climate Data to the World" at the DOE Climate Change Modeling Program Science Team Meeting in Gaithersburg, MD, March 2010. The conference was held in Gaithersburg, MD, March 29–April 2, 2010.

5.1.5   *NOAA Climate Observations and Models Workshop*

Dean N. Williams participated in the presentation titled, "Observation for CMIP-5 Simulations" given at the Climate Observation and Models Workshop in Asheville, NC, March 24, 2010.

5.1.6   *High Performance Computing in Weather and Climate Conference*

Wang Feiyi presented, "Petascale Data Management in Support of DOE and NSF Communities" at the High Performance Computing in Weather and Climate in Beijing, China March 22, 2010.

5.1.7 *WMO Information Systems Meeting, Inter-Asia Workshop on WIS*

- The Japanese Meteorological Agency (JMA) hosted an Inter-Asia and Middle East workshop on building the WMO WIS system March 9-12, 2010 in Tokyo, Japan. As an invited lecturer, Don Middleton presented an overview of NCAR's participation in WMO WIS, which included the federation of ESG-CET capabilities.

5.1.8 *GlobusWorld 2010*

- Rachana Ananthakrishnan presented, "Earth System Grid: Community Update", at the GlobusWorld 2010 conference held at the Argonne Illinois Conference Center, March 2–4, 2010.

- Frank Siebenlist presented, "Earth System Grid Data Access Security", at the GlobusWorld 2010 conference held at the Argonne Illinois Conference Center, March 2–4, 2010.

5.1.9 *Ocean Sciences Meeting*

- Steve Hankin et al. presented NOAA efforts to compliment and leverage the work of the ESG in a talk titled, "The Unified Access Framework for Gridded Data" at the 2010 Ocean Sciences Meeting in Portland, Oregon, February 22–26, 2010 (http://www.agu.org/meetings/os10/).

- Jeremy Malczyk and Steve Hankin presented ESG software components in a talk titled, "Using Software Version Control Tools for Data Management and Quality Control Collaboration in the Surface Ocean CO2 Atlas (SOCAT) Project" at the 2010 Ocean Sciences Meeting in Portland, Oregon, February 22–26, 2010 (http://tinyurl.com/SOCAT-AGU-OS-2010).

5.1.10 *WMO Information Systems Meeting, Intercommission Coordination Group (WMO ICG-WIS)*

- Don Middleton represented NCAR and ESG at the WMO ICG-WIS meeting in Seoul, Korea February 22-26, 2010. His presentation was entitled "NCAR DCPC Update: With a Focus on CMIP5/IPCC, and on Behalf of an International Collaboration".

5.1.11 *Oak Ridge National Laboratory Meetings*

- Galen Shipman presented, "The Earth System Grid at ORNL" to representative of the National Climatic Data Center, October 21, 2009.

- Galen Shipman presented, "The Earth System Grid" to representatives of the University of Idaho, February 1, 2010.

- Galen Shipman presented, "Facilitating a Data Intensive Computational Environment at the OLCF" to representatives of the Exxon Corporation, February 17, 2010.

5.1.12 *LBNL Magellan and Advanced Network Initiative Meeting*

Dean N. Williams and Alex Sim presented, "The Earth System Grid Center for Enabling Technologies: Network and Analysis Requirements" to the leads of ESnet and LBNL Magellan projects. The meeting took place at LBNL in Berkeley, California, November 10, 2009.

5.1.13 *NOAA Global Interoperability Program Kickoff Meeting*

Luca Cinquini, Steve Hankin, Don Middleton, and Dean N. Williams presented the "Earth System Grid Center for Enabling Technologies" at the NOAA funded Global Interoperability Program (GIP) kickoff meeting held at the Geophysical Fluid Dynamics Laboratory in Princeton, NJ on November 5–6, 2009.

5.1.14 *2009 Global Organization for Earth System Science Portal (GO-ESSP) Workshop*

At the GO-ESSP workshop, the following presentations were presented by ESG-CET team members: Luca Cinquini – "An Introduction to the ESG-CET CMIP-5 Federation", Dean N. Williams and Bob Drach – "ESG Data Node Configuration", Dean N. Williams and Karl Taylor – "CMIP-5 Overview", and "Climate Model Output Rewriter", Steve Hankin – "Report from OceanObs09: Opportunities and challenges for the emerging framework of netCDF-CF-THREDDS-DAP" and "Remote Processing Issues", Roland Schweitzer – "Model Intercomparison Using the LAS Interactive Earth Science Data Visualization Galary (vizGal)", and Gary Strand – "Distributing Climate Model Data to the World: A 10 year retrospective". The workshop was held October 6–8 at the Institute for Pharmacy in Hamburg, Germany.

Posters:

5.1.15 *DOE BER Climate Change Modeling Program Science Team Meeting*

Dean N. Williams and Galen Shipman presented a poster titled, "The Earth System Grid Center for Enabling Technologies (ESG-CET): Serving Climate Data to the World" at the DOE Climate Change Modeling Program Science Team Meeting in Gaithersburg, MD, March 2010.

5.1.16 *SC'09 Conference*

Brady, E., Chervenak, A. L., "Mirroring Earth System Grid Data Sets", Supercomputing (SC09) Conference, ACM Student Poster Session, Portland, Oregon, November 2009.